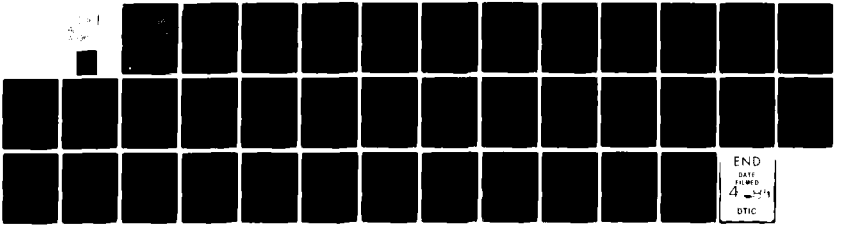
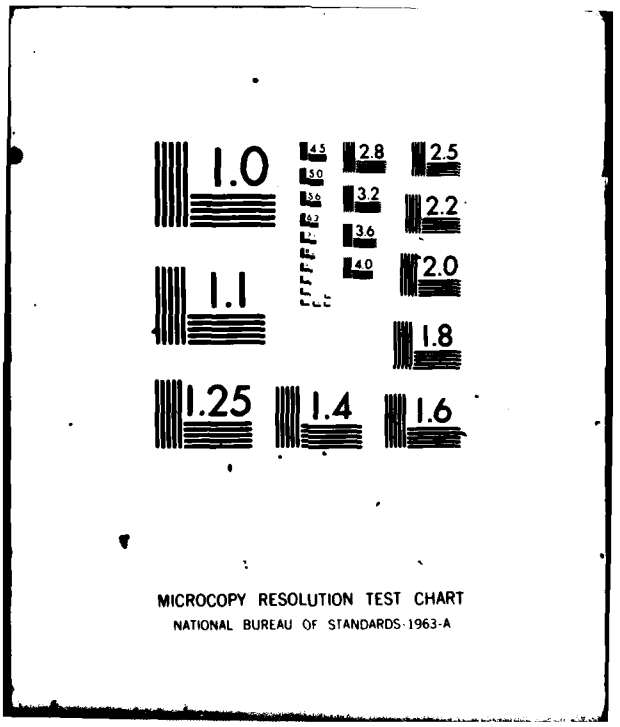
AD-A096 188 NORTHWESTERN UNIV EVANSTON ILL DEPT OF INDUSTRIAL E--ETC F/G 12/1
MODELING THE DISTRIBUTION OF FINGERPRINT CHARACTERISTICS, REVIS--ETC(U)
SEP 80 S L SCLOVE N00014-80-C-0408
UNCLASSIFIED DISCUSSION PAPER-50 NL



END
DATE FILMED
4-81
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
10.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

NORTHWESTERN UNIVERSITY

# Center for Statistics and Probability

LEVEL II



Discussion Paper Number 50

September, 1980

## MODELING THE DISTRIBUTION OF FINGERPRINT CHARACTERISTICS

Stanley L. Sclove

Department of Industrial Engineering and Management Sciences

Northwestern University

Evanston, Illinois 60201

# (6) MODELING THE DISTRIBUTION OF FINGERPRINT CHARACTERISTICS. Revision 1.

by

(10) STANLEY L. SCLOVE

(21) Presented at the

NATO ADVANCED STUDY INSTITUTE/INTERNATIONAL SUMMER SCHOOL ON STATISTICAL DISTRIBUTIONS IN SCIENTIFIC WORK,

Trieste, Italy,

10-31 July [redacted] 1980

(14) Discussion Paper-50, TR-80-2



(9) TECHNICAL REPORT NO. 80-2

(11) 19 Sep[redacted]80

(12) 38

PREPARED FOR THE
OFFICE OF NAVAL RESEARCH
(15) UNDER
CONTRACT N00014-80-C-0408
TASK NR042-443

Principal Investigator: Stanley L. Sclove



CENTER FOR PROBABILITY AND STATISTICS
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS 60201

rev. 1 (9/19/80)

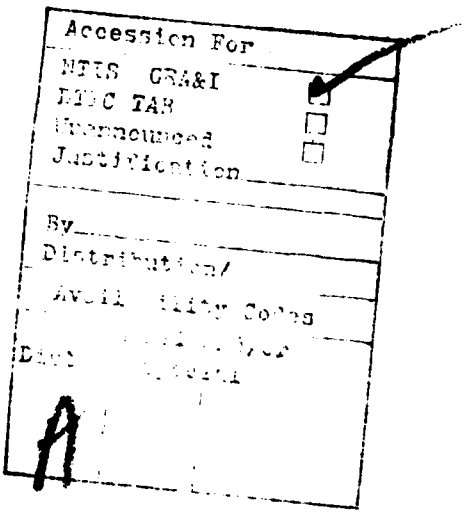
# MODELING THE DISTRIBUTION OF FINGERPRINT CHARACTERISTICS

STANLEY L. SCLOVE
Northwestern University and University of Illinois at Chicago Circle

Department of Industrial Engineering and Management Sciences
The Technological Institute
Northwestern University
Evanston, Illinois 60201, U.S.A.

CONTENTS





# MODELING THE DISTRIBUTION OF FINGERPRINT CHARACTERISTICS

STANLEY L. SCLOVE
Northwestern University and University of Illinois at Chicago Circle

Department of Industrial Engineering and Management Sciences
The Technological Institute
Northwestern University
Evanston, Illinois 60201, U.S.A.

## ABSTRACT

Quantitative aspects of fingerprints are discussed. A study undertaken to develop methods for assigning probabilities to partial fingerprints is summarized, with emphasis on distributional aspects.

## KEYWORDS

fingerprints; two-way series; multinomial distribution; Markov process; Poisson process

## 1. INTRODUCTION

This paper focuses on the distributional aspects of a study reported in three earlier articles--Osterburg, Parthasarathy, Raghavan, Sclove (1977), Sclove (1979), Sclove (1980a)--concerning the assignment of probabilities to partial fingerprints based on the numbers and locations of occurrences of the ten Galton characteristics. In the study a grid of cells was superimposed on the fingerprints. The number of characteristics in the cells is modeled as a multivariate two-way series (i.e., a multivariate stochastic process with two-dimensional indexing parameter). The statistical parameters were estimated from the data (fingerprints). Estimation of the probability of partial prints is illustrated. Some comparisons are made with the estimates provided by an assumption of independence between cells. Some analysis based on statistical results for infinitely divisible distributions ~~[see Sclove (1980b)] is~~ discussed.

## 2. BACKGROUND INFORMATION ON FINGERPRINTS

### 2.1. Types

The three types. The bulb of each finger of the human hand contains ridge lines that form themselves into patterns, thus providing a basis for classification. Ridge-line patterns are of three major types: loops (ca. 65%), whorls (ca. 30%), and arches (ca. 5%). There is further subdivision within each major pattern. Arches are either plain or tented, loops are radial or ulnar, whorls are plain, central pocket loops, double loops, or accidentals. This further subdivision within each pattern allows a classification scheme to be organized so that for the ten fingers many categories of fingerprint-pattern combinations result. Within each category there are many fingerprints from different individuals which, to the untrained eye, appear to be the same. This process of separation through classification results in relatively small sets of fingerprints which are of manageable proportions for the purpose of search and comparison.

Search. Chernoff (1977) has treated the problem of selecting a subset of files such as fingerprint files for careful comparison with a target print to decide if the corresponding individual is represented in the files. It is assumed that much of the data in the files and on the target are subject to noise or random error. The (likelihood-ratio) solution depends upon the joint distribution of the filed data and the target data and their marginal distributions.

Computer assistance. Computer classification of single fingerprints into types (subdivisions of arch, loop, and whorl) by a syntactic approach has been achieved; see, e.g., Rao and Balck (1980).

Enhancement of latent fingerprints by numerical processing of the image has been treated; see, e.g., Chiralo and Berdan (1978).

## 2.2. Ridge counts

In the loop pattern there is a point where the three opposing ridge systems come together. (The outer and the lower ridge lines change concavity at that point.) This point is the triradius, or delta. If a straight line is drawn from the delta to the core, a certain number of ridge lines will be crossed. This number is the ridge-count. Patterns with no triradius (simple arches) have no ridge count. In the case of patterns with two triradii (whorls and double loops) there are two counts; sometimes then one just works with the higher count. Sometimes the sum across fingers of the individual ridge counts, when defined, is considered; it is also called the "ridge count."

Holt (1951-2) has studied the correlation between numbers of crossings on different fingers.

## 2.3. The Galton details

Fingerprints and dermatoglyphics in general have found use in medicine and genetics as correlates of genetic abnormalities; see, e.g., Holt (1968) and Priest, Tishler and Rosner (1976). The emphasis here, however, is on the use of fingerprints in identification, as in criminalistics. Partial prints such as those left at crime scenes do not always permit determination of the type or number of crossings. Even if they did, the individuality of a print would have to be based on the details of the print. The ridge-line details are termed Galton characteristics since Sir Francis Galton was among the first to study them systematically [Galton (1892)]. He defined ten kinds of minutiae. One is a ridge ending, an abrupt ending to a ridge line; ridge endings are by far the most frequent characteristic. A ridge line may suddenly divide into two branches, much like a fork in a road; such a characteristic is termed a bifurcation (or fork). Similarly, eight other characteristics are defined. There is general agreement upon these ten types of ridge-line details. [See, e.g., Osterburg, Parthasarathy, Raghavan and Sclove (1977) for details and giagrams of the ten characteristics. See Appendix A for working definitions used for some of the characteristics.] The purpose of the study discussed here was to model the occurrence of these Galton characteristics, with a view toward the development of formulas for

the calculation of probabilities of partial fingerprints.

The study was made as follows. A grid of one millimeter squares was placed over a fingerprint. Each fingerprint is considered as a configuration of the cells of the grid. For each cell of the grid there are several possibilities: one or more of the ten characteristics is there, or no characteristic is present. Thus a configuratrion is a grid of cells, where each cell may be thought of as either being empty or else being occupied by one or more words, the words representing the characteristics present. E.g., if a cell contains the words "dot, dot, ending ridge," it means that the area corresponding to that cell contained two dots and one ending ridge. Table 1 shows a configuration of 43 cells, with 4 ridge endings and two forks.

[INSERT TABLE 1.]

A match between a suspect's full print and a partial print exists when there is a section of the full print that is the same as the partial print. Since we are working in terms of a grid of cells, for our purposes a match exists when a grid can be laid on the full print in such a way that the resulting configuration contains a section which is the same as the configuration corresponding to the partial print.

The fingerprints studied were enlarged to ten times actual size, making a full rolled print about 8" by 10". The cells of the grid were one centimeter square after enlargement. Members of the project staff coded the ten Galton characteristics, cell by cell. (See Appendix A for precise working definitions of the characteristics.) Thirty-nine prints were coded. (Osterburg had earlier examined 40 prints, from 40 different individuals, but one was missing, leaving 39 for re-examination.) There is no problem with representativeness of the sample. The Galton characteristics are "accidental." They are not genetic. With regard to these characteristics, two siblings, even two twins, are no more alike than two random persons.

[On this point see e.g., Kingston (1964, p. 26), and the references given there.] Therefore, with respect to the Galton characteristics, each and every person is "representative."

## 3. DATA DESCRIPTION

By an occurrence we mean the occurrence of any one of the ten Galton characteristics. The 39 fingerprints used yielded a total of 8591 cells which could be coded. In all there were 2536 occurrences, or 0.295 per cell. Table 2 gives the distribution of the number of occurrences per cell, without regard to type.

[INSERT TABLE 2]

The abbreviations used for the characteristics are as follows.

B: bridge

D: dot

E: ending ridge

F: fork (bifurcation)

I: island

L: lake (eye)

O: delta

S: spur

T: trifurcation

Z: double bifurcation

The symbol DE, for example, denotes the occurrence of one dot and one ending ridge in a cell; BEE would denote the occurrence of a bridge and two ending ridges in a cell; etc. Altogether 54 combinations occurred, including one DEEEE and one DDDDE. Table 3 gives the distribution of these cell configurations. (Of course, with a larger data set, many more cell configurations would occur.) Note in particular that 77% of the cells were empty; i.e., the probability that a cell is occupied is .23.

[INSERT TABLE 3]

## 3.1. Data processing

One physical record corresponded to one cell and took the following form. (The abbreviation "cc." means "card columns.")

| cc. 1-2 | cc.3 | cc.4 | cc.5-6 | cc. 7-8 | cc. 9-13 |
|---|---|---|---|---|---|
| Fingerprint number, n | Hand | Finger | Row, i | Column, j | Alphabetic information giving cell contents |
| 25 | R | I | 13 | 11 | BEE |

The line of data above signifies that there is a bridge (B) and two ridge endings (E) in the cell corresponding to row 13 and column 11 of fingerprint number 25, which is from the index finger (I in cc. 4) of somebody's right hand (R in cc. 3).

Actually these data were not card-punched but rather typed on a terminal and stored directly on disk, so "cc." is used only figuratively. Cc. 9-13 contain alphabetic information giving the contents of the cell. This shows the need either for programming in a language such as PL1 which allows alphabetic variables or for use of a text editor to convert the alphabetic data to numerical. The latter method was used, the field of cc. 9-13 being replaced by a field of ten columns (cc. 9-18) of the form X(1), X(2), ... , X(10), where, for $v = 1, 2, \ldots, 10$, X(v) is the number of occurrences of the v-th characteristic in the cell. E.g., BEE would be translated as (0,1,0,0,2,0,0,0,0,0) since, in the numbering used for the ten characteristics, X(2) = number of bridges and X(5) = number of ridge endings.

## 3.2. Notation

The process of occurrence of the Galton characteristics was modeled as a multivariate two-dimensional stochastic process, more specifically, a ten-variate process with two-dimensional indexing parameter. The index designates location (row and column) in the grid. The ten variates are the numbers of occurrences of the ten Galton characteristics. That is, the process is $\{\underset{\sim}{X}_{ij}, (i,j) \text{ in } G\}$, where G

is the set of cells corresponding to the fingerprint impression. If the impression were rectangular, with I rows and J columns, then the grid G would be simply {(i,j): i = 1, 2,..., I, j = 1, 2,..., J}. Let the subscript n range over the 39 prints. Then the data set is

$$\{x_{\sim nij}, \; n = 1, 2, \ldots, 39, \; (i,j) \text{ in } G_n\},$$

where $G_n$ denotes the grid of usable cells in the n-th print. The basic scalar datum is $x_{vnij}$, the number of occurrences of the v-th Galton characteristic (v=1,2,...,10) in the (i,j)-th cell of the n-th print. Note that (i,j) is nested in n, in the sense that (i,j) has no absolute meaning; it is not the case that the core (center) of the print always has the same location.

## 4. THE MULTINOMIAL MODEL

Two aspects of the modeling process are modeling within cells and modeling between cells. Osterburg, Parthasarathy, Raghavan, Sclove (1977) used a multinomial model within cells and independence between cells. Sclove (1979) used the same multinomial model within cells but a Markov model between cells. Sclove (1980a) used the same Markov model between cells but a Poisson model within cells.

This section treats the multinomial model. The next section treats probabilities of various configurations under the multinomial model with independence. Section 6 summarizes the multinomial model with a Markov between-cells model. Section 7 summarizes the Poisson within-cells model in the context of the Markov between-cells model.

The multinomial within-cells model is as follows: For any cell there are 13 possibilities; either the cell is empty, or one of the following twelve possibilities has occurred: B, D, E, F, I, L, O, S, T, Z, EE (broken ridge), or other multiple occurrence. (By "multiple occurrence" we mean more than one occurrence in a cell.)

In regard to the selection of the multinomial categories priority was given

to the ten standard Galton characteristics, occurring as singletons. The number of possible combinations (multiple occurrences) of these individual characteristics is enormous. Among the combinations, we selected the double ridge ending because it was the most frequent; also, it includes a broken ridge, which is different from a ridge coming to an end. A consequence of lumping rare multiple occurrences together into the single category "other multiple occurrence" is to give the benefit of the doubt to the suspect, in the sense of giving a conservative, i.e., large, probability estimate for the given configuration.

In terms of random variables the use of the multinomial model corresponds to using random vectors

$$\underset{\sim nij}{Y} = (Y_{0nij}, Y_{1nij}, \ldots, Y_{12,nij}),$$

defined as follows.

$Y_{0nij}$, indicator of empty cell $= 1$ if $\underset{\sim nij}{X} = (0,0,0,0,0,0,0,0,0,0)'$
$= 0$ otherwise

$Y_{1nij}$, indicator of island $= 1$ if $\underset{\sim nij}{X} = (1,0,0,0,0,0,0,0,0,0)'$
$= 0$ otherwise

$Y_{2nij}$, indicator of bridge $= 1$ if $\underset{\sim nij}{X} = (0,1,0,0,0,0,0,0,0,0)'$
$= 0$ otherwise

.
.
.

$Y_{10,nij}$, indicator of delta $= 1$ if $\underset{\sim nij}{X} = (0,0,0,0,0,0,0,0,0,1)'$
$= 0$ otherwise

$Y_{11,nij}$, indicator of two ridge endings $= 1$ if $\underset{\sim nij}{X} = (0,0,0,0,2,0,0,0,0,0)'$
$= 0$ otherwise

$Y_{12,nij}$, indicator of multiple occurrence $= 1$ if $Y_{vnij} = 0$ for $v = 0,1,2,\ldots,11$
$= 0$ otherwise

## 5. THE MULTINOMIAL MODEL WITH INDEPENDENCE

The model we employ in this section can be summarized as follows. First there are the two within-cells modeling assumptions developed in the preceding section.

(1) A fingerprint is characterized as a configuration of the cells of a grid.

(2) For any cell there are 13 possibilities; either the cell is empty, or one of the following twelve possibilities has occurred: B, D, E, F, I, L, O, S, T, Z, EE (broken ridge), or other multiple occurrence. (By "multiple occurrence" we mean more than one occurrence in a cell.)

Now, for between-cells modeling, we consider

(3) The cells are statistically independent.

Assumption (1) is used throughout the study; (2) is used in this and the next section but replaced in Section 7 by a Poisson assumption; (3) is used in this section and replaced in Section 6 by a Markov model.

The probability P of a given configuration is, under this model, given by the point multinomial distribution, as

$$\log P = k(0)\log P(0) + k(1)\log P(1) + \ldots + k(12)\log P(12),$$

where the k(i), i = 0, 1, 2, ..., 12, are non-negative integers summing to t, the total number of cells in the print, and the P(i)'s are the probabilities of the 13 possibilities and hence sum to one. (For notational reasons and because the probabilities involved are small it is convenient to work in terms of logarithms.)

Estimates of probabilities. The parameters P(i) of the model were estimated from the data. See Table 4. The variance of the estimate of any one of the P(i) is P(i)[1-P(i)]/n, i = 0, 1, 2, ... , 12, where n = 8591 cells. Table 4 gives estimates p(i) of P(i), i = 0, 1, 2, ...,

12, and also estimates of the corresponding standard deviations. (We use upper case P for the parameter and lower case p for the estimate.)

[INSERT TABLE 4.]

The probability P of a configuration of k(0) empty cells, k(1) cells containing islands, k(2) cells containing bridges, ..., k(10) cells containing deltas, k(11) cells containing two ending ridges, and k(12) cells containing other multiple occurrences is estimated by an estimate p given by

$$\log p = k(0)\log p(0) + k(1)\log p(1) + \ldots + k(12)\log p(12).$$

Let E, for entropy (information) be defined as $E = -\log P$ and $e = -\log p$. We have

$$E = -k(0)\log P(0) - k(1)\log P(1) - \ldots - k(12)\log P(12).$$

Appendix B gives confidence bounds for E, based on the estimate e.

The study of inter-cell dependence discussed in the next section [from Sclove (1979)] indicates that the approximations of the present section should give results that are sufficiently accurate.

The preceding has dealt with the assignment of a probability P to the occurrence of a given configuration in a given set of cells. For inferential purposes it is necessary to estimate the probability that a person has this configuration anywhere on his fingers. A discussion of this aspect of the problem is given in Osterburg, Parthasarathy, Raghavan and Sclove (1977).

## 6. THE MULTINOMIAL MARKOV MODEL

The next analysis, relating to dependence among cells, shows that the probability that a cell is occupied increases monotonically with the number of neighbors occupied. Square blocks of 9 cells, 3 cells by 3 cells, were examined to determine the extent of inter-cell dependence. The data set of Osterburg, Parthasarathy, Raghavan and Sclove (1977) yielded 845 such blocks of cells. For i = 1, 2, ..., 845 blocks, let

the variable y(i) = 1 or 0 according as the center cell of the i-th block is occupied or not, and let x(i) be the number of adjacent cells which are occupied; x(i) is between 0 and 8. Table 5 gives the cross-tabulation of y and x and gives, for each value of x, the proportion of y's that are equal to 1, i.e., the proportion of center cells which are occupied.

[INSERT TABLE 5.]

The probability of occupancy increases monotonically with x. Such absolute consistency was not expected, firstly because it seems so rare in data analyses and secondly because it was thought that occurrences in most of the adjacent cells might crowd out occurrence in the center cell.

The value of the chi-square statistic for testing independence based on Table 5 is 18.77 ($P<.005$, 6 d.f., the categories x = 6,7, and 8 having been pooled). The decomposition of this overall value based on the value .1404 of the correlation coefficient between x and y is given in Table 6. The value of chi-square due to correlation, 16.65, is the sample size (845 blocks) times the square of the correlation coefficient.

[INSERT TABLE 6.]

[In order to achieve independent trials for the chi-square test, separate blocks of 9 cells were used. This greatly reduces the effective sample size. The results here were clear, so it was not necessary to be more efficient. It should be noted, however, that such problems can be handled in a more efficient manner by Besag's (1974) "coding" scheme, used and discussed in later sections where necessary.]

The above analysis demonstrated the necessity of developing a model which took account of inter-cell dependence. (It was subsequently found that the model based on inter-cell independence gave adequate results, but this determination could be made only in the context of a model incorporating dependence.) Accordingly, then, the model of Section 5 was extended to consider inter-cell dependence and the occurrence of the characteristics was modeled as a two-dimensional Markov-

type process. This analysis is reported in Sclove (1979). The model can be described by saying that it is a nearest-neighbor, Markov-type model where the conditioning is on the sum and the allocation across types of characteristics is independent of the value of the sum.

Under this model, the estimated probability of the configuration of Table 1 is -12.0 . Compare this with the figure of -11.4 given by the approximation based on an assumption of independence between cells. The difference in logarithms is 0.6; the ratio of the two estimates is thus 4:1. This difference is unimportant since we are interested only in order of magnitude. Note further that the estimate based on independence is a larger probability, i.e., it is conservative in this sense. [See Osterburg, Parthasarathy, Raghavan and Sclove (1977) for some discussion of the bearing of these probabilities on the guilt or innocence of a suspect. A large probability estimate is conservative in favor of a suspect, in the sense that it gives the suspect the benefit of the doubt.] In general, independence gives too much weight (too low a probability) to configurations with a lot of clustering of occurrences. In the configuration of Table 1 there is some but not a great deal of clustering.

## 7. THE POISSON MARKOV MODEL

The categories defined in Assumption 2 are somewhat arbitrary. The ten categories corresponding to the occurrence of each of the ten characteristics as singletons are natural enough; it is the lumping together of multiple occurrences which warrants alternative treatment. In the preceding section the occurrence of the characteristics was modeled as a two-dimensional multinomial process, taking account of dependence among cells but not dealing differently with the problem of multiple occurrences. In Sclove (1980a) the occurrence of the characteristics is modeled as a two-dimensional Poisson process, not only taking account of dependence among cells but also providing alternative treatment of multiple occurrences.

According to the between-cells data analysis discussed above, the probability that a cell is occupied increases monotonically with the number of neighbors occupied. Accordingly, we introduced an assumption that the expected number of occurrences in a cell depends upon the outcomes in neighboring cells only through the number of such cells that are occupied.

A within-cells data analysis is discussed in Appendix C. It was found that negative binomial distributions provided a good fit to the distribution of the number of characteristics per cell, and to the numbers of different characteristics. This is consistent with a model of a mixture of Poisson distributions, for a negative binomial distribution can be obtained as a mixture of Poisson distributions. Accordingly, we set out to test the hypothesis that the number of occurrences in a given cell is a Poisson random variable, at least conditionally.

These assumptions combined into an assumption that the number of occurrences in a cell is distributed according to a Poisson distribution with parameter $M$, say, which depends upon the random variable $A$, the number of adjacent cells occupied. In other words, the conditional distribution of the number of occurrences, given the number of adjacent cells that are occupied, i.e., given $A = a$, is Poisson with parameter $M(a)$, $a = 0, 1, 2,$

3, or 4.

This assumption was tested by fitting the number of occurrences of characteristics for each fixed number of adjacencies to a Poisson distribution and checking the goodness of fit. In making this test the dependence among cells had to be taken into account. The problem of dependence was treated by a method of "coding" discussed by Besag (1974); see the discussion by Bartlett (1975), p. 27. To understand the method, suppose the cells were labelled as in Table 7 with two symbols, y and o. This

[INSERT TABLE 7]

allows the values at the y-sites to be taken, conditional on the values at the o-sites, as independent. Table 8 gives the number of occurrences, by number of adjacencies, for the "y" cells in Besag's coding scheme. The results for the four orientations are given. The generalized likelihood ratio test was used to compare the Poisson fit with the empirical distribution. The "chi-square" values in Table 8 are values of -2 ln L, where L is the generalized likelihood ratio. The Poisson fit appeared adequate. (The Pearson chi-square gave similar results.) Accordingly, a model was developed, based on these assumptions. Details are given in Sclove (1980a).

[INSERT TABLE 8]

This Poisson Markov model gave for the configuration of Table 1 an estimated log probability of -12.0. Compare this with result of -11.4 given by the approximation based on independence and -11.8 given by the multinomial Markov model. We have 12.0 - 11.4 = 0.6; the ratio of the two corresponding estimates is about 4:1.

## 8. THE INFINITELY DIVISIBLE MODEL

Alternative models considered include modeling the observed random vector giving the numbers of the ten characteristics per cell as an infinitely divisible random vector. [See Sclove (1980b) for a

discussion of multivariate infinitely divisible random vectors.]

A random variable $X$ (which may be a scalar, vector or matrix) is infinitely divisible if there exists a triangular sequence $Y(1,1)$; $Y(2,1)$, $Y(2,2)$; ...; $Y(n,1)$, $Y(n,2)$, ..., $Y(n,n)$; ..., such that, for each $n = 1, 2, \ldots,$ the n random variables $Y(n,1)$, $Y(n,2)$, ..., $Y(n,n)$ are independent and identically distributed and the variables $X(1)$, $X(2)$, ..., $X(n)$, ..., defined by

$$X(1) = Y(1,1)$$

$$X(2) = Y(2,1) + Y(2,2)$$

$$\vdots$$

$$X(n) = Y(n,1) + Y(n,2) + \ldots + Y(n,n)$$

$$\vdots$$

all have the same distribution as $X$.

An assumption that $\underset{\sim}{X} = (X_1, X_2, \ldots, X_{10})'$ in the fingerprint study is infinitely divisible can be supported on both physical and probabilistic grounds, as follows.

Speaking first from the physical point of view, it is not at all unreasonable to consider a point process to be infinitely divisible. For, the random variables count the numbers of occurrences in some specified area, such as the cells of the grid. One can conceive of using finer and finer grids. The Y's in the decomposition necessary for infinite divisibility correspond to the cells of these finer partitions. The Galton characteristics may be considered as occurring at dimensionless points, a fork occurring at the point of bifurcation, a spur at the point of separation, and so on. Thus the assumption of infinite divisibility seems reasonable.

Arguing from probabilistic grounds, the assumption of infinite divisibility, at least under a hypothesis of independence of the variates, seems justified, on the grounds that negative binomial and

Poisson distributions, which fit the marginal distributions, are infinitely divisible.

As discussed in Sclove (1980b), Pierre (1971) defines the measure of dependence (h here, pi in his notation)

$$h(X,Y) = \mathrm{Cov}[(X-EX)^2,(Y-EY)^2] - 2[\mathrm{Cov}(X,Y)]^2$$

for random variables X,Y in an infinitely divisible random vector with no Gaussian component. He further shows that $[h(X,Y)/[h(X,X)h(Y,Y)]^{1/2}$ is between zero and one and hence is a normalized measure of dependence analogous to a correlation coefficient. [The parameter h(X,Y) is the cumulant of order (2,2) of (X,Y); the corresponding k-statistic estimate can be used.] Estimates of h were used to estimate the normalized measure for the 10 x 9/2 = 45 pairs of characteristics. The values were small; in fact, the largest was only .018. (The square root of this is still only .13.) Thus an assumption of independence of the ten variates is further supported by this analysis.

## ACKNOWLEDGEMENTS

The initial study, reported in Osterburg, Parthasarathy, Raghavan, and Sclove (1977), was supported under a contract with the Center for Research in Criminal Justice, University of Illinois at Chicago Circle.

Work on Sclove (1979) and Sclove (1980a) was supported under Grant AFOSR 77-3454 from the Air Force Office of Scientific Research. Computations were performed using the facilities of the Computer Center of the University of Illinois at Chicago Circle.

The author's current work on Markov models for two-way series is supported by the Office of Naval Research as Contract N00014-80-C-0408 under Task NR 042-443. These sources of support are gratefully acknowledged.

## REFERENCES

Bartlett, M. S. (1975). The Statistical Analysis of Spatial Pattern, in the series (formerly Methuen's) Monographs on Applied Probability and Statistics. Chapman and Hall, London; Halsted Press, New York.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B, 36, 192-236.

Bowman, K. O., Hutcheson, K., Odum, E. P., and Shenton, L. R. (1971). Comments on the distribution of indices of diversity. In Statistical Ecology, Volume 3: Many Species Populations, Ecosystems, and Systems Analysis, G. P. Patil, E. C. Pielou, and W. E. Waters, eds. Pennsylvania State University Press, University Park, Pennsylvania.

Chernoff, H. (1977). Some applications of a method of identifying an element of a large multidimensional population. In Multivariate Analysis-IV, P. R. Krishnaiah, ed. North-Holland Publishing Company, Amsterdam (Elsevier North-Holland, New York).

Chiralo, R. P., and Berdan, L. L. (1978). Adaptive digital enhancement of latent fingerprints. Proceedings of the Society of Photo-Optical Instrumentation Engineers, 149, 118-125.

Cooke, T. D. (1974). Personal communication to Osterburg.

Cox, D. R., and Miller, H. D. (1965). The Theory of Stochastic Processes. Wiley, New York.

Galton, F. (1892). Finger Prints. Macmillan, London; republication (1965), DaCapo Press, New York.

Greenwood, M., and Yule, G. U. (1920). An inquiry into the nature of frequency-distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. Journal of the Royal Statistical Society, 83, 255.

Holt, S. (1951-2). The correlations between ridge-counts on different fingers. Annals of Eugenics, 16, 287-297.

Holt, S. (1968). The Genetics of Dermal Ridges. Charles C. Thomas, Springfield, Ill.

Kingston, C. R. (1964). Probabilistic analysis of partial fingerprint patterns. Doctoral dissertation, University of California, Berkeley.

Kingston, C. R. (1965a). Applications of probability theory in criminalistics--I. Journal of the American Statistical Association, 60, 70-80.

Kingston, C. R. (1965b). Applications of probability theory in criminalistics--II. Journal of the American Statistical Association, 60, 1028-1034.

Osterburg, J. W., Parthasarathy, T., Raghavan, T. E. S., and Sclove, S. L. (1977). Development of a mathematical formula for the calculation of fingerprint probabilities based on individual characteristics. Journal of the American Statistical Association, 72, 772-778.

Parzen, E. (1962). Stochastic Processes. Holden-Day, Inc., San Francisco.

Pierre, P. A. (1971). Infinitely divisible distributions, conditions for independence, and central limit theorems. Journal of Mathematical Analysis and Applications, 33, 341-354.

Priest, J. H., Tishler, P. V., and Rosner, B. (1976). Dermatoglyphics in mosaic Down's syndrome. Clinical Genetics, 9, 417-426.

Rao, K., and Balck, K. (1980). Type classification of fingerprints: a syntactic approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2, 223-231.

Sclove, S. L. (1979). The occurrence of fingerprint characteristics as a two-dimensional process. Journal of the American Statistical Association, 74, 588-595.

Sclove, S. L. (1980a). The occurrence of fingerprint characteristics as a two-dimensional Poisson process. Communications in Statistics, A9, 675-695.

Sclove, S. L. (1980b). Some recent statistical results for infinitely divisible distributions. Proceedings of the International Summer School/NATO Advanced Study Institute on Statistical Distributions in Scientific Work, Trieste, Italy, July 10-31, 1980, G. P. Patil and C. Taillie, eds.

## APPENDIX A: The Galton Characteristics

Definitions of some of the Galton characteristics were refined by means of precise working definitions, necessary to accomplish the coding.

A bridge was defined as less than two centimeters in length in the enlarged photograph (i.e., two millimeters in actuality); otherwise, it would be coded as a fork.

A dot was defined as being large enough to encompass one pore. Smaller "dots" were not counted; larger "dots" were coded as short ridges.

Distinct breaks in ridges were coded as two separate ending ridges to distinguish such breaks from ridges simply coming to an end.

A spur was defined as being less than two centimeters in length in the enlargement (i.e., two millimeters in actuality); otherwise, it was coded as a fork. A spur was counted only once: the end of a spur was not counted as a ridge ending.

The sizes used are of an order suggested by T. Dickerson Cooke of the Institute of Applied Science, Chicago, Illinois [Cooke (1974)] and are consistent with recommendations of the Committee on Standardization of the International Association for Identification.

APPENDIX B: Derivation of Confidence Bounds for the Entropy

The negative log probabilities considered in the model based on independence are in terms of logs base 10 and are given by the expression

$$E = -\log_2 P = -[k(0) \log P(0) + k(1) \log P(1) + \ldots + k(10) \log P(10)],$$

where $P(0)=1-P(1)-P(2)-\ldots-P(12)$ and $k(0)=t-k(1)-k(2)-\ldots-k(12)$, t being the total number of cells in the print. For the estimate e of E we have $e = -\log p = c \ln p = cH$, where $H = \ln p$ and the constant c is the log base 10 of the base "e" of the natural logs (about 0.434). Thus $Var(e) = c\ Var(H)$. The asymptotic variance of H is given by Bowman et al. (1971); this gives

$$Var(e) \approx (1/n)c^2 \sum_{i=0}^{12} [k(i)]^2/P(i) - t^2].$$

This variance is estimated by substituting the estimates p(i) for the P(i). E.g., for 12 ending ridges and no other characteristics in a print of area t = 72 cells, we have 60 empty cells. Hence k(0) = 60, k(5) = 12, and the other k(i) are 0. From Table 4 we have p(0) = .766 and p(5) = .0832. Thus $e = -12 \log_2 .0832 - 60 \log_2 .766 = 19.9$, and $Var(e) \approx (1/8591)\ 0.434\ [(60^2/.766) + (12^2/.0832) - 72^2] = 0.0273$. The corresponding standard deviation, the square root of this, is 0.165. Thus a 95% confidence interval is obtained from the point estimate by adding and subtracting 1.96(0.165) = 0.3.

## APPENDIX C: Marginal Distribution of the Characteristics

The distribution of the number of occurrences per cell is given in Table 2.

For testing goodness-of-fit, it was necessary to use Besag's coding scheme to achieve independent trials. This gave the distribution of Table 9. A Poisson distribution is inadequate [$-2 \ln L = 9.69$, 3 d.f., $P = .021$; the Pearson chi-square gave a similar result: chi-square statistic = 8.63, 3 d.f. (pooling categories), $P = .03$]. The distribution is well fit by a negative binomial distribution; in fact, the special case of a geometric distribution provides an adequate fit (Pearson chi-square = 3.11, 3 d.f., $P = .38$).

[INSERT TABLE 9]

An interpretation of the fit by the negative binomial family is that what is involved is a gamma-type mixture of Poisson distributions [see, e.g., Parzen (1962, p. 57)], resulting in a negative binomial distribution, as in the classical accident studies of Greenwood and Yule (1920). Empirical support for the assumption is demonstrated by the plausibility of the following assumptions [the usual axioms for a Poisson process [see, e.g., Parzen (1962, p. 119)], generalized to two dimensions]. Given any set S in the (x,y)-plane, let N(S) be the number of occurrences in S and let a(S) be the area of S. Given any point (x,y), let {S(n)} be a sequence of sets tending to (x,y) as n tends to infinity. Then the following assumptions are plausible. There is a positive number M(x,y) such that, as n tends to infinity,

$\{1 - Pr[N(S(n))=0]\}/a(S(n))$ tends to $M(x,y)$,

$Pr[N(S(n))=2]/a(S(n))$ tends to $M(x,y)$,

$Pr[N(S(n))\geq 2]/a(S(n))$ tends to 0 .

The intensity parameter varies with position in the sense that occurrences are more probable in the pattern than the non-pattern area and the intensity may be a decreasing function of the distance from the core of the pattern. The two-dimensional non-homogeneous Poisson process may be

termed a Poisson random field, since the intensity parameter varies with position. See, e.g., Cox and Miller (1965) for a discussion of such Poisson processes over general spaces. The monograph by Bartlett (1975) provides a discussion of general models and methods for analysis of two-dimensional processes; see the Appendix in Sclove (1979) for a synopsis.

LIST OF CAPTIONS FOR TABLES

1. Configuration of 43 cells with 4 ending ridges and 2 forks.

   0 = empty cell, E = ending ridge, F = fork.

2. Distribution of number of occurrences

3. Distribution of cell configurations

4. Estimates of probability parameters

5. Cross-tabulation of occupancy of center cell

   and number of adjacent cells occupied

6. Decomposition of chi-square according to correlation between x and y

7. Coding scheme for obtaining conditionally-independent trials

   in a second-order process

8. Distribution of number of occurrences, by number of adjacencies,

   with Besag's coding scheme, for test of goodness-of-fit

   of Poisson distribution

9. Distribution of number of occurrences

   for subsample of independent cells

1. Configuration of 43 cells with 4 ending ridges and 2 forks. 0 = empty cell, E = ending ridge, F = fork.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| a | 0 | 0 | 0 | 0 | 0 | 0 |
| b | E | 0 | 0 | 0 | E | 0 |
| c | 0 | 0 | F | 0 | 0 | 0 |
| d | 0 | 0 | 0 | 0 | 0 | 0 |
| e | 0 | 0 | 0 | E | 0 | 0 |
| f | 0 | 0 | 0 | 0 | F | 0 |
| g | 0 | E | 0 | 0 | 0 | 0 |
| h | | | 0 | | | |

## 2. Distribution of number of occurrences

| Number of occurrences | 0 | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| Number of cells | 6584 | 1594 | 320 | 72 | 19 | 2 | 8591 |
| Proportion of cells | .766 | .185 | .0372 | .00838 | .0022 | .0023 | 1.00 |

3. Distribution of cell configurations

| Cell configuration | Frequency | |
|---|---|---|
| | Number of cells | Percent of cells |
| Empty | 6584 | 76.6% |
| E | 715 | 8.32 |
| F | 328 | 3.82 |
| I | 152 | 1.77 |
| D | 130 | 1.51 |
| EE | 119 | 1.39 |
| B | 105 | 1.22 |
| S | 64 | 0.745 |
| L | 55 | 0.640 |
| EL | 32 | 0.372 |
| DE | 32 | 0.372 |
| EEE | 21 | 0.244 |
| EI | 21 | 0.244 |
| O | 17 | 0.198 |
| DD | 15 | 0.175 |
| BE | 13 | 0.151 |
| Z | 12 | 0.140 |
| DI | 11 | 0.128 |
| EEEE | 10 | 0.116 |
| ES | 10 | 0.116 |
| DDI | 10 | 0.116 |
| II | 9 | 0.105 |
| FI | 9 | 0.105 |
| BF | 7 | 0.0815 |
| DEE | 7 | 0.0815 |
| FF | 5 | 0.0582 |
| T | 5 | 0.0582 |
| EEF | 4 | 0.0466 |
| BEE | 4 | 0.0466 |
| EII | 4 | 0.0466 |
| FL | 3 | 0.0349 |
| BB | 3 | 0.0349 |
| FS | 2 | 0.0233 |
| BD | 2 | 0.0233 |
| DDE | 2 | 0.0233 |
| LL | 2 | 0.0233 |
| Other (19 other multiple occurrences) | 67 | 0.780 |
| Total | 8591 cells | 100.0% |

## 4. Estimates of probability parameters

| Probability parameter | Cell configuration | Frequency | Estimate of probability parameter | Estimated standard deviation of estimate |
|---|---|---|---|---|
| P(0) | Empty | 6584 | .766 | .0045 |
| P(1) | Island (I) | 152 | .0177 | .0014 |
| P(2) | Bridge (B) | 105 | .0122 | .0012 |
| P(3) | Spur (S) | 64 | .00745 | .00093 |
| P(4) | Dot (D) | 130 | .0151 | .0013 |
| P(5) | Ending ridge (E) | 715 | .0832 | .0030 |
| P(6) | Fork (F) | 328 | .0382 | .0021 |
| P(7) | Lake (L) | 55 | .00640 | .00086 |
| P(8) | Trifurcation (T) | 5 | .000582 | .00024 |
| P(9) | Double bifurcation (Z) | 12 | .00140 | .00040 |
| P(10) | Delta (O) | 17 | .00198 | .00048 |
| P(11) | Broken ridge (or EE) | 118 | .0139 | .0013 |
| P(12) | Other multiple occurrence | 305 | .0355 | .0020 |
| | | 8591 | 1.0 | |

Source: Table 3

5. Cross-tabulation of

occupancy of center cell

and number of adjacent cells occupied

| x | y | | |
|---|---|---|---|
| | 0 | 1 | Total |
| 0 | 152(84.4) | 28( 15.6) | 180(100) |
| 1 | 170(79.15) | 45( 20.9) | 215(100) |
| 2 | 163(78.4) | 45( 21.6) | 208(100) |
| 3 | 97(77.0) | 29( 23.0) | 126(100) |
| 4 | 44(65.7) | 23( 34.3) | 67(100) |
| 5 | 23(63.9) | 13( 36.1) | 36(100) |
| 6 | 7(58.3) | 5( 41.7) | 12(100) |
| 7 | 0(---) | 0(----) | 0(---) |
| 8 | 0( 0.0) | 1(100.0) | 1(100) |
| | 656 | 189 | 845 blocks of cells |

y = 1 if given (center) cell is occupied

= 0 if it is empty

x = number of adjacencies (number of adjacent cells occupied)

6. Decomposition of chi-square

according to correlation between x and y

| Source of Variation | d.f. | Value of Chi-square | |
|---|---|---|---|
| Overall | 6 | 18.77 | (P<.005) |
| Correlation | 1 | 16.65 | (P<.005) |
| Residual | 5 | 2.12 | (.80<P<.85) |

7. Coding scheme for obtaining conditionally-independent trials in a second-order process

o o o o o o o o

o y o y o y o y

o o o o o o o o

o y o y o y o y

8. Distribution of number of occurrences, by number of adjacencies, with Besag's coding scheme, for test of goodness-of-fit of Poisson distribution

| No. of adjacencies | Orientation | Number of occurrences 0 | 1 | 2 | 3 | 4 | n | Mean | Variance | Chi-square | d.f. | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 174 | 34 | 7 | 1 | 0 | 216 | .24 | .27 | 2.79 | 2 | .25 |
| | 2 | 175 | 34 | 4 | 0 | 0 | 220 | .20 | .20 | 0.58 | 1 | .45 |
| | 3 | 180 | 37 | 2 | 0 | 0 | 219 | .19 | .17 | 1.19 | 1 | .28 |
| | 4 | 187 | 41 | 7 | 0 | 0 | 235 | .23 | .24 | 1.64 | 1 | .20 |
| 1 | 1 | 161 | 44 | 5 | 1 | 0 | 211 | .27 | .27 | 0.55 | 2 | .76 |
| | 2 | 177 | 32 | 8 | 0 | 3 | 220 | .27 | .44 | 26.9 | 3 | <.001 |
| | 3 | 167 | 47 | 8 | 1 | 2 | 225 | .33 | .43 | 10.1 | 3 | .02 |
| | 4 | 152 | 46 | 6 | 2 | 0 | 206 | .31 | .33 | 1.80 | 2 | .41 |
| 2 | 1 | 78 | 33 | 7 | 1 | 1 | 120 | .45 | .52 | 2.54 | 3 | .47 |
| | 2 | 76 | 25 | 6 | 3 | 0 | 110 | .42 | .52 | 4.48 | 2 | .11 |
| | 3 | 70 | 28 | 8 | 0 | 0 | 106 | .42 | .40 | 2.48 | 1 | .11 |
| | 4 | 77 | 1 | 7 | 3 | 0 | 98 | .35 | .56 | 17.3 | 2 | <.001 |
| 3 | 1 | 24 | 12 | 5 | 2 | 1 | 44 | .73 | .99 | 2.88 | 3 | .41 |
| | 2 | 25 | 10 | 3 | 0 | 1 | 39 | .51 | .73 | 5.06 | 3 | .17 |
| | 3 | 21 | 8 | 4 | 2 | 0 | 35 | .63 | .83 | 3.27 | 2 | .19 |
| | 4 | 23 | 15 | 2 | 1 | 0 | 41 | .54 | .50 | 1.48 | 2 | .48 |
| 4 | 1 | 4 | 2 | 1 | 0 | 0 | 7 | .57 | .62 | 0.48 | 1 | .48 |
| | 2 | 3 | 5 | 2 | 0 | 0 | 10 | .90 | .54 | 2.08 | 1 | .15 |
| | 3 | 7 | 2 | 3 | 0 | 0 | 12 | .67 | .79 | 3.62 | 1 | .06 |
| | 4 | 2 | 4 | 3 | 0 | 0 | 9 | 1.11 | .61 | 2.96 | 1 | .08 |

n's differ somewhat due to border effects; the grids are not perfect rectangles.

9. Distribution of number of occurrences

for subsample of independent cells

| | Number of occurrences | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | n | Mean |
| Number of cells | 441 | 125 | 25 | 5 | 2 | 0 | 598 | 0.331 |
| Proportion of cells | .737 | .209 | .042 | .008 | .003 | .000 | 1.000 | |

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|
| 1. REPORT NUMBER: Technical Report 80-2 | 2. GOVT ACCESSION NO.: AD-A096 188 |
| 3. RECIPIENT'S CATALOG NUMBER: | |
| 4. TITLE (and Subtitle): Modeling the Distribution of Fingerprint Characteristics | 5. TYPE OF REPORT & PERIOD COVERED: Technical Report |
| | 6. PERFORMING ORG. REPORT NUMBER: |
| 7. AUTHOR(s): Stanley L. Sclove | 8. CONTRACT OR GRANT NUMBER(s): N00014-80-C-0408 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Mathematics, University of Illinois at Chicago Circle, Box 4348, Chicago, IL 60680 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: |
| 11. CONTROLLING OFFICE NAME AND ADDRESS: — | 12. REPORT DATE: September 19,1980 |
| | 13. NUMBER OF PAGES: 34 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Research, Statistics and Probability Branch, Arlington, VA | 15. SECURITY CLASS. (of this report): Unclassified |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: |

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
Unlimited distribution

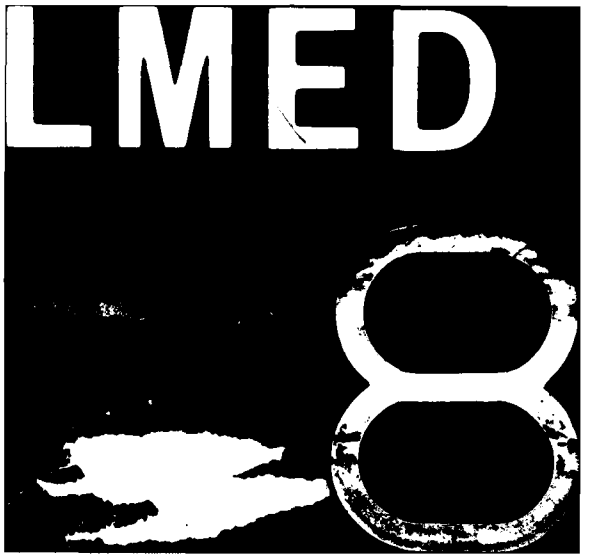
18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
fingerprints; two-way series; multinomial distribution; Markov process; Poisson process

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Quantitative aspects of fingerprints are discussed. A study undertaken to develop methods for assigning probabilities to partial fingerprints is summarized, with emphasis on distributional aspects.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


LMED
8